4.5 Estimation of Average Effects

In causal inference problems, one can often categorize methods for estimating treatment effects as being based on regression, weighting, or both (doubly robust).

For the average treatment effect $\psi = \mathbb{E}(Y^1 - Y^0)$, regression estimators can be motivated based on the identifying expression

$$\psi = \mathbb{E}\Big\{\mathbb{E}(Y^1 \mid X, A=1) - \mathbb{E}(Y^0 \mid X, A=0)\Big\} = \mathbb{E}\Big\{\mu_1(X) - \mu_0(X)\Big\}$$

which suggests the regression estimator

$$\widehat{\psi}_{reg} = \mathbb{P}_n\Big\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\Big\}. \tag{4.2}$$

Operationally, this estimator predicts (the conditional mean of) the potential outcomes Y^1 and Y^0 for each subject, takes the difference, and averages across the sample. One can also interpret this estimator with reference to matching: for each unit with a particular $X = x$ value, one finds unit(s) with the same or similar $X = x$ value but who received the opposite treatment.

Weighting estimators can be motivated based on the inverse-probability-weighted expression

$$\psi = \mathbb{E}\left[\left\{\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right\}Y\right] = \mathbb{E}\Big\{\mu_1(X) - \mu_0(X)\Big\}$$

which suggests the inverse-probability-weighted estimator

$$\widehat{\psi}_{ipw} = \mathbb{P}_n\left[\left\{\frac{A}{\widehat{\pi}(X)} - \frac{1-A}{1-\widehat{\pi}(X)}\right\}Y\right]. \tag{4.3}$$

This estimator can be viewed as up- or down-weighting observations whose covariates are under- or over-represented in their treated group compared to the population covariate distribution. This is similar in spirit to importance sampling: the covariate distribution for the treated is different from that in the general population, so one needs to use a change of measure to reweight treated outcomes appropriately. It is also popular to view the inverse weighting as creating a "pseudopopulation" of treated units whose covariate distribution matches that of the entire population.

Doubly robust estimators can be motivated based on the expressions

$$\begin{aligned}\psi &= \mathbb{E}\left[\left\{\frac{A}{\overline{\pi}(X)} - \frac{1-A}{1-\overline{\pi}(X)}\right\}\Big\{Y - \mu_A(X)\Big\} + \Big\{\mu_1(X) - \mu_0(X)\Big\}\right] \\ &= \mathbb{E}\left[\left\{\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right\}\Big\{Y - \overline{\mu}_A(X)\Big\} + \Big\{\overline{\mu}_1(X) - \overline{\mu}_0(X)\Big\}\right]\end{aligned}$$

which hold for any $(\overline{\pi}, \overline{\mu})$. This suggests the estimator

$$\widehat{\psi}_{dr} = \mathbb{P}_n \left[\left\{ \frac{A}{\widehat{\pi}(X)} - \frac{1-A}{1-\widehat{\pi}(X)} \right\} \Big\{ Y - \widehat{\mu}_A(X) \Big\} + \Big\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \Big\} \right] \tag{4.4}$$

which was also used in the previous chapter with experiments, except now the propensity score $\pi(x)$ depends on covariates and is unknown so needs to be estimated. The doubly robust estimator is somewhat less intuitive than the other two options, but it can be viewed as correcting leftover smoothing bias of a regression of inverse-probability-weighted estimator, or augmenting an inverse-probability-weighted estimator with regression predictions to increase efficiency. We will see in later chapters that its precise form comes from a bias correction based on a distributional Taylor expansion of the average treatment effect functional.

4.5.1 Discrete Covariates

For some intuition we will first consider the simplest case, where the covariates X are discrete and low-dimensional, i.e., $X \in \{1, ..., d\}$ with d fixed. We will see that in this setup, when one uses the empirical distribution to estimate the "nuisance functions" π and μ_a, then all three of the previously mentioned estimators coincide in that they are numerically equivalent. (Later we will show that they are asymptotically efficient in a local minimax sense). This numerical equivalence does not occur when the covariates have some continuous components and modeling or smoothing is used to construct the $\widehat{\pi}$ and $\widehat{\mu}_a$ estimates. Intuitively, the reason why all three estimators are numerically equivalent is because, when the covariates are discrete, there is no smoothness or additional structure to exploit, so each estimator is making full equivalent use of the data. Another way to think about it is that, in the discrete case, the empirical measure $\mathbb{P}_n$ is an actual valid distribution (including all conditional distributions), and so the identifying expression equalities above also hold for $\mathbb{P}_n$.

Our first result shows the numerical equivalence between the regression, weighting, and doubly robust estimators.

Proposition 4.4. *Suppose $X \in \{1, ..., d\}$ is discrete and the nuisance estimators are the empirical averages*

$$\widehat{\pi}(x) = \mathbb{P}_n(A \mid X = x) = \frac{\mathbb{P}_n\{A\mathbb{1}(X = x)\}}{\mathbb{P}_n\{\mathbb{1}(X = x)\}}$$
$$\widehat{\mu}_a(x) = \mathbb{P}_n(Y \mid X = x, A = a) = \frac{\mathbb{P}_n\{Y\mathbb{1}(A = a)\mathbb{1}(X = x)\}}{\mathbb{P}_n\{\mathbb{1}(A = a)\mathbb{1}(X = x)\}}$$

Then the regression, weighting, and doubly robust estimators defined in (4.2)–(4.4) are all numerically equivalent, i.e.,

$$\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}.$$

Proof. We will consider the $\psi_1 = \mathbb{E}\{\mu_1(X)\}$ term, since the logic is the same for ψ_0. To see that $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw}$ note that

$$\begin{aligned}
\widehat{\psi}_{reg} &= \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{P}_n\{YA\mathbb{1}(X = x_i)\}}{\mathbb{P}_n\{A\mathbb{1}(X = x_i)\}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{1}{n}\sum_j Y_j A_j \mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_i)\}} = \frac{1}{n}\sum_{j=1}^{n} Y_j A_j \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n} Y_j A_j \frac{\frac{1}{n}\sum_i \mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_j)} = \frac{1}{n}\sum_{j=1}^{n} Y_j A_j / \widehat{\pi}(X_j) = \mathbb{P}_n\left\{\frac{AY}{\widehat{\pi}(X)}\right\} = \widehat{\psi}_{ipw}
\end{aligned}$$

where in the fifth equality we replace the x_i in the denominator with x_j since the numerator includes the indicator $\mathbb{1}(X_j = x_i)$.

Now to see that $\widehat{\psi}_{reg} = \widehat{\psi}_{dr}$ we will show $\mathbb{P}_n\{AY/\widehat{\pi}(X)\} = \mathbb{P}_n\{A\widehat{\mu}_1(X)/\widehat{\pi}(X)\}$, so that the correction term $\mathbb{P}_n[A\{Y - \widehat{\mu}_1(X)\}/\widehat{\pi}(X)] = 0$. Note

$$\begin{aligned}
\mathbb{P}_n\left\{\frac{A\widehat{\mu}_1(X)}{\widehat{\pi}(X)}\right\} &= \frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{\widehat{\pi}(X_i)} \frac{\frac{1}{n}\sum_j Y_j A_j \mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n} Y_j A_j \frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{\widehat{\pi}(X_i)} \frac{\mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n} Y_j A_j \frac{1}{\widehat{\pi}(X_j)} \frac{1}{n}\sum_{i=1}^{n} A_i \frac{\mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_j)} \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{Y_j A_j}{\widehat{\pi}(X_j)} = \mathbb{P}_n\left\{\frac{AY}{\widehat{\pi}(X)}\right\}
\end{aligned}$$

where again in the third equality we replace x_i in the denominator with x_j due to the numerator indicator. This gives the result. □

Next we derive the limiting distribution of the estimator $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}$.

Theorem 4.1. *Suppose $X \in \{1, ..., d\}$ is discrete and the nuisance estimators are the empirical averages from Proposition 4.4. Assume that Y is bounded and that $\pi(x)$ and $\widehat{\pi}(x)$ are bounded away from ϵ and $1-\epsilon$ for some $\epsilon > 0$ and all x. Then*

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N(0, var(f))$$

for $\widehat{\psi}$ the estimators in (4.2)–(4.4) and

$$f(Z) = \mu_1(X) - \mu_0(X) + \left\{\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right\}\Big\{Y - \mu_A(X)\Big\}.$$

Proof. We will work with the $\widehat{\psi}_{dr}$ version of the estimator, which can be written as $\widehat{\psi}_{dr} = \mathbb{P}_n(\widehat{f})$ for

$$f(Z) = \mu_1(X) - \mu_0(X) + \left\{ \frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)} \right\} \Big\{ Y - \mu_A(X) \Big\}$$

and $\widehat{f}$ the version of f replacing (π, μ_a) with $(\widehat{\pi}, \widehat{\mu}_a)$.

Therefore by Lemma 3.1 we have the decomposition

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})f + (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) + \mathbb{P}(\widehat{f} - f) \equiv Z^* + T_1 + T_2.$$

We will first handle the T_1 term. Note since X is discrete we can write the nuisance estimators $(\widehat{\pi}, \widehat{\mu}_a)$ as linear regression estimators based on saturated models, i.e.,

$$\widehat{\pi}(x) = \pi(x; \widehat{\alpha}) = \widehat{\alpha}^{\mathrm{T}} w$$

where $w^{\mathrm{T}} = \{\mathbb{1}(x=1), ..., \mathbb{1}(x=d-1)\} \in \{0,1\}^{d-1}$ and similarly

$$\widehat{\mu}_a(x) = \mu_a(x; \widehat{\beta}_a) = \widehat{\beta}_a^{\mathrm{T}} w.$$

This implies

$$|\widehat{f}(z) - f(z)| = |f(z; \widehat{\eta}) - f(z; \eta)| \leq C \|\widehat{\eta} - \eta\|$$

for $\eta = (\alpha, \beta_0, \beta_1)$ and $C < \infty$ some constant. Therefore f and $\widehat{f}$ belong to a Donsker class, which together with the central limit theorem and Lemma 19.24 from van der Vaart [2000] imply that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$.

For the T_2 term, note that $f = f_1 - f_0$ for $f_a = \mu_a + \frac{\mathbb{1}(A=a)(Y-\mu_a)}{a\pi(x)+(1-a)\{1-\pi(x)\}}$. Then

$$\begin{aligned}
\mathbb{P}(\widehat{f}_1 - f_1) &= \mathbb{P}\left[\frac{A}{\widehat{\pi}(X)} \Big\{ Y - \widehat{\mu}_1(X) \Big\} + \Big\{ \widehat{\mu}_1(X) - \mu_1(X) \Big\} \right] \\
&= \mathbb{P}\left[\frac{\pi(X)}{\widehat{\pi}(X)} \Big\{ \mu_1(X) - \widehat{\mu}_1(X) \Big\} + \Big\{ \widehat{\mu}_1(X) - \mu_1(X) \Big\} \right] \\
&= \mathbb{P}\left[\frac{\pi(X) - \widehat{\pi}(X)}{\widehat{\pi}(X)} \Big\{ \mu_1(X) - \widehat{\mu}_1(X) \Big\} \right] \\
&\leq \mathbb{P}\left\{ \left| \frac{\pi(X) - \widehat{\pi}(X)}{\widehat{\pi}(X)} \right| \Big| \mu_1(X) - \widehat{\mu}_1(X) \Big| \right\} \\
&\leq \left(\frac{1}{\epsilon} \right) \mathbb{P}\left\{ \Big| \pi(X) - \widehat{\pi}(X) \Big| \Big| \mu_1(X) - \widehat{\mu}_1(X) \Big| \right\} \\
&\leq \left(\frac{1}{\epsilon} \right) \|\pi - \widehat{\pi}\| \|\mu_1 - \widehat{\mu}_1\| \\
&= O_{\mathbb{P}}(1/\sqrt{n}) O_{\mathbb{P}}(1/\sqrt{n}) = O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(1/\sqrt{n})
\end{aligned}$$

where the second and third lines used iterated expectation, the fifth used the bound on $\widehat{\pi}$, the sixth used Cauchy-Schwarz, and the last line used that $\widehat{\pi}$ and $\widehat{\mu}_a$ are root-n consistent due to the discrete (e.g., they can be represented as linear regression estimators, as mentioned above). The same exact logic follows for $\mathbb{P}(\widehat{f}_0 - f_0)$, which then yields the result since $T_1 + T_2 = o_{\mathbb{P}}(1/\sqrt{n})$. □

Theorem 4.1 shows that, when the covariates are discrete and low-dimensional, the causal effect estimators $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}$ are all root-n consistent and asymptotically normal under only mild boundedness conditions. The key to proving this result was the analysis of the T_2 term; the logic used there will be repeated throughout the book going forward.

Theorem 4.1 gives confidence intervals (and thus hypothesis tests) as an immediate corollary.

Corollary 4.1. *Under the conditions of 4.1, an asymptotically valid confidence interval for the average treatment effect ψ is given by*

$$\widehat{\psi} \pm 1.96\sqrt{\widehat{var}(\widehat{f})/n}.$$

Remark 4.4. Although the regression, weighting, and doubly robust estimators are exactly equal, to construct confidence intervals we need to estimate the asymptotic variance with the empirical variance of the terms appearing in the doubly robust estimator.

In summary, when the measured covariates are sufficient to control confounding, and are discrete and low-dimensional, the choice of estimator is immaterial – regression, weighting, and doubly robust estimation are all numerically equivalent and efficient. In the next section, however, we will see that the story is much different in the more realistic scenario where the covariates are not discrete and some modeling is necessary.

Appendix A

Notation Guide

Y^a	Potential outcome under treatment/exposure $A = a$
$\perp\!\!\!\perp$	Statistically independent
$\xrightarrow{p}$	Convergence in probability
$\rightsquigarrow$	Convergence in distribution
$O_{\mathbb{P}}(1)$	Bounded in probability
$o_{\mathbb{P}}(1)$	Converging in probability to zero
$\mathbb{P}_n$	Sample average operator, as in $\mathbb{P}_n(\widehat{f}) = \mathbb{P}_n\{\widehat{f}(Z)\} = \frac{1}{n}\sum_{i=1}^n \widehat{f}(Z_i)$
$\mathbb{P}$	Conditional expectation given the sample operator, as in $\mathbb{P}(\widehat{f}) = \int \widehat{f}(z)\ d\mathbb{P}(z)$
$\lVert \cdot \rVert$	$L_2(\mathbb{P})$ norm $\lVert f \rVert = \sqrt{\mathbb{P}(f^2)}$ or Euclidean norm, depending on context
$\lVert \cdot \rVert_1$	$L_1(\mathbb{P})$ norm $\lVert f \rVert_1 = \mathbb{P}(\lvert f \rvert)$
$\lVert \cdot \rVert_\infty$	L_∞ or supremum norm $\lVert f \rVert_\infty = \sup_z \lvert f(z) \rvert$
$\mathcal{H}(s)$	Hölder class of functions with smoothness index s
$\lesssim$	Less than or equal, up to a constant multiplier

Bibliography

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1993.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20(3):261, 2005.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

S. Leon, A. A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055, 2003.

D. Michaels. *Doubt is their product: how industry's assault on science threatens your health*. Oxford University Press, 2008.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13. Springer, 1982.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

J. M. Robins and A. Rotnitzky. Comments on: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.

J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1): 113–124, 2000.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

D. B. Rubin and M. J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.

L. Yang and A. A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.